Naive Bayes

Hoàng Vũ Minh

Naive Bayes dựa trên định lý Bayes:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Đây là công thức tính xác suất có điều kiện của A khi xuất hiện B, trong đó:

- P(A|B) is Probability of A given B, the probability of A given that B happens
- P(B\A) is Probability of B given A, the probability of B given that A happens
- P(A), P(B) lần lượt là xác suất xảy ra A và B

Naive Bayes Classifier:

Thuật toán này có từ NAIVE (ngây thơ) vì nó xuất phát từ một giả định ngây thơ: rằng mọi features đầu vào đều độc lập với nhau. Điều này rất khó xảy ra trong thực tế. Vì các feature là độc lập nên ta có công thức tính xác suất:

$$p(\mathbf{x}|c) = p(x_1, x_2, \ldots, x_d|c) = \prod_{i=1}^{d} p(x_i|c)$$

Ưu điểm: Nhờ vào tính ngây thơ -> công thức tính toán nhanh, test nhanh
Nhược điểm: Độ hiệu quả có thể không được đảm bảo

Công thức xác định class của biến x mới:

$$c = \arg \max_{c \in \{1,\ldots,C\}} p(c) \prod_{i=1}^{d} p(x_i|c)$$

p(x) dưới mẫu được loại bỏ vì nó không phụ thuộc vào c

- trong đó: p(c) được tính bằng cách lấy số lần lớp đó xuất hiện chia cho tổng số các lớp xuất hiện, là xác suất tiên nghiệm.
- p(xi/c) là xác suất đặc trưng x xuất hiện ở lớp C, tính bằng cách đếm số lần xuất hiện của x trong lớp c chia cho tổng số đặc trưng trong lớp đó

Tức là khi đưa một biến mới vào, ta sẽ tính tích xác suất có điều kiện của điểm dữ liệu mới, sau đó tính xác suất cuối cùng và chọn ra class có xác suất cao nhất làm predict.

Có ba biến thể chính của Naive Bayes classifier tùy thuộc vào dạng của dữ liệu:

- Bernoulli Naive Bayes: Dùng cho dữ liệu nhị phân, mỗi đặc trưng có hai giá trị có thể (0 hoặc 1).
- Multinomial Naive Bayes: Dùng cho dữ liệu rời rạc, thường được sử dụng trong phân loại văn bản với đặc trưng là tần suất từ, có thể kết hợp với BOW (Bag of words) để thực hiện một vài bài toán NLP đơn giản.
- Gaussian Naive Bayes: Dùng cho dữ liệu liên tục, giả định rằng các đặc trưng tuân theo phân phối chuẩn (Gaussian).

Multinomial Naive Bayes:

Chủ yếu được dùng trong phân loại văn bản, với dữ liệu đầu vào là một Bag of Words. Bag of Words là một thuật ngữ trong lĩnh vực xử lý ngôn ngữ tự nhiên, giúp chuyển hóa các văn bản thành các vector chứa thông tin về tần suất xuất hiện của từ đó trong văn bản.

Ưu điểm: đơn giản, dễ hiểu, dễ tính toán, nhược điểm: mất ngữ cảnh.

$$\lambda_{ci} = p(x_i|c) = \frac{N_{ci}}{N_c}$$

- p(xi/c) tỉ lệ với tuần suất từ thứ i xuất hiện trong các văn bản của class c
- Nci là tổng số lần từ thứ i xuất hiện trong các văn bản của class
- c, nó được tính là tổng của tất cả các thành phần thứ i của các feature vector tương ứng với class c.

Gaussian Naive Bayes

Sử dụng trong trường hợp dữ liệu là các biến liên tục, tuân theo phân phối Gaussian

$$p(x_i|c) = p(x_i|\mu_{ci}, \sigma_{ci}^2) = \frac{1}{\sqrt{2\pi\sigma_{ci}^2}} \exp\left(-\frac{(x_i - \mu_{ci})^2}{2\sigma_{ci}^2}\right)$$

Bernoulli Naive Bayes:

Mô hình này được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị binary - bằng 0 hoặc 1. Ví dụ, thay vì đầu vào là bag of word như trên kia, ta có thể chỉ cần tạo một vector chứa thông tin một từ nào đó có xuất hiện hay không.

Ta vẫn có thể dùng Bernoulli Naive Bayes cho bài toán phân loại văn bản như trên Multinomial, nhưng cần chuyển dữ liệu từ tần suất thành nó xuất hiện (1) hay không xuất hiện (0).

Khi đó, $p(x_i|c)$ được tính bằng:

$$p(x_i|c) = p(i|c)^{x_i}(1 - p(i|c)^{1-x_i}$$

với $p(i|c)$ có thể được hiểu là xác suất từ thứ i xuất hiện trong các văn bản của class c.